HCMUS_Beyonders at AstroCLIMB: Same Paper or Citing Paper? Linking Astronomy Figures with OCR, Citation Graphs, and LLMs

My-Linh Bang[1,2,†] **and** **Phu-Quy Nguyen-Lam**[1,2,†] **and** **Duy-Minh Dao-Sy**[1,2]
Trung-Kiet Huynh[1,2] and **Phu-Hoa Pham**[1,2] and **Chi-Nguyen Tran**[1,2]
[1]Faculty of Information Technology, University of Science (HCMUS), Vietnam
[2]Vietnam National University - Ho Chi Minh City (VNU-HCM), Vietnam
{23122009, 23122048}@student.hcmus.edu.vn
{23122041, 23122039, 23122030, 23122044}@student.hcmus.edu.vn
[†]Equal contribution

Abstract

AstroCLIMB asks whether two objects taken from astronomy papers, each a figure or a caption, show the same figure, come from the same paper, come from papers linked by a citation, or are unrelated. We describe the two-stage system of team HCMUS_Beyonders, which reached a macro-F1 of 0.7421 on the full test set. In the first stage, cross-encoders read both objects of a pair jointly. Each figure is represented by the text that optical character recognition (OCR) extracts from it, an input that scored higher than every image-embedding and image-patch input we tried. Because test papers are unseen, the cross-encoders are pretrained on 600K pairs generated from the citation graph of the public AstroCLIMB corpus. An error analysis shows that this stage most often confuses pairs from the same paper with pairs from citing papers, and that none of the cues we tested separates the two. The second stage therefore adds a different kind of model rather than more features: a Qwen3-8B large language model (LLM) fine-tuned with low-rank adaptation (LoRA). On its own it scores lower than the first stage, but over a third of each stage's errors are not made by the other. Averaging the two raises held-out macro-F1 by 0.016 (95% confidence interval $[+0.002, +0.030]$) and test macro-F1 from 0.7072 to 0.7329. Adding a second LLM, Gemma-4-12B, without held-out validation brings the test score to 0.7421, a gain within test noise.

1 Introduction

AstroCLIMB (Grezes et al., 2026), which follows earlier shared tasks of this workshop series (Grezes et al., 2022, 2023, 2025), asks a system to relate two objects taken from astronomy papers, where an object is a figure image or a figure caption. The system must decide whether the two objects are the same figure (*same_figure*), come from the same paper (*same_paper*), come from two papers connected by a citation (*related_papers*), or are unrelated (*unrelated_papers*). Such links support figure-level search in the NASA Astrophysics Data System, which released the data.

The task is difficult for three reasons. First, most astronomy figures are plots, so encoders trained on photographs capture their style but not their content, which lives in axis labels, object names, and legends. Second, only 10,000 labelled pairs exist, and the test pairs come from papers absent from both the training set and the public AstroCLIMB corpus (§2), so a model must generalise to unseen papers. Finally, a paper and the papers it cites usually share a topic, and a pair carries no title, author, or identifier that would separate them.

We address the first two difficulties in a first stage built on text-centred cross-encoders, models that read the two objects of a pair as one input sequence. Instead of embedding figure pixels, we read the text inside each figure with optical character recognition (OCR), and a BERT-style model (Devlin et al., 2019) reads it together with the other object. Before training on the task data, we pretrain these models on 600K labelled pairs generated from the citation graph of the public corpus.

An error analysis of this first stage shows that the remaining gap is the third difficulty: it most often confuses *same_paper* with *related_papers*, failing on *same_paper* pairs that look different and *related_papers* pairs that look alike. None of the further cues we tested separates them, including object names, cross-references, and a contrastive objective that removes topic. The second stage therefore adds a different kind of model instead of another feature: large language models (LLMs) fine-tuned with low-rank adaptation (LoRA). Their errors only partly coincide with those of the first stage; we do not attribute this to their size (Appendix G).

Our contributions are: (1) evidence that OCR text represents astronomy figures better than the image embeddings and image patches we tried,

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 52–60
November 10, 2026 ©2026 Association for Computational Linguistics

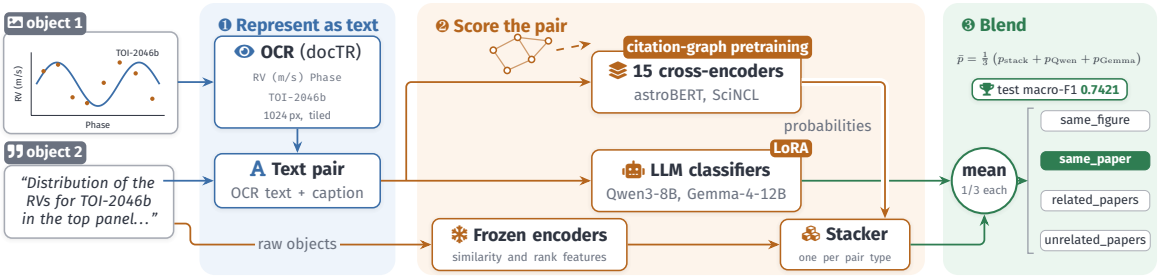


Figure 1: System overview. Images enter the text models as OCR text. In the first stage, cross-encoder probabilities and similarities from frozen (not fine-tuned) encoders feed a stacker; the second stage averages its output with two LLMs. The label is the argmax, with *same_figure* allowed only for image–caption pairs.

which did not include a vision–language LLM; (2) citation-graph pretraining, which raised test macro-F1 by 0.022; (3) an error analysis that locates the dominant error, and a blend with two LoRA-tuned LLMs that raises test macro-F1 from 0.7072 to 0.7421. Our code is available at `https://github.com/NLPQuy/WASP_2026`. Related work is discussed in Appendix I.

2 Task and Data

Each of the 10,000 training and 10,000 test pairs has one of three *pair types*: image–image, image–caption, or caption–caption. *same_figure* can only hold for image–caption pairs, which leaves ten allowed combinations of pair type and label; each has exactly 1,000 training pairs. Systems are ranked by macro-F1 over the four classes on the full test set. The leaderboard shows this *test score* after every submission, with no hidden split (Appendix D).

Test papers appear neither in the training set nor in the *public corpus* (Grezes et al., 2026), the only extra data we use: a release of 94,233 figures with their captions, paper DOIs, and the reference and citation lists of each paper. Test objects cannot be looked up there: only 2 of 5,256 test captions match a public caption, whereas all 9,325 training captions do, and 0.5% of test images match a public image by perceptual hash, against 38.5% of training images.

3 System

Figure 1 shows the system, which we describe in the order we built it. The first stage is the *cross-encoder system* (§3.1–3.3); the second averages its output with two LLM classifiers (§3.4), and we call this average the *blend*. Because the relation is symmetric, every text model reads both orders of a pair and averages the two predictions.

3.1 Representing figures as text

We read each figure with docTR (Mindee, 2021), using a DB-ResNet50 detector (Liao et al., 2020) and a PARSeq recogniser (Bautista and Atienza, 2022). The image is resized to 1024 pixels and split into tiles so that small tick labels are recovered. An image object is then represented by its OCR text and a caption object by its caption, so every pair is a pair of texts.

On an early version of the system, a cross-encoder that read only text reached 0.556 cross-validated macro-F1. Every model that received the image itself reached at most 0.533, whether as a pooled embedding or patch tokens from SigLIP 2 (Tschannen et al., 2025) or the chart encoder MatCha (Liu et al., 2023) (Appendix E). We did not test a vision–language LLM that reads the raw figure, so this finding is limited to those image inputs. Reading figures at 1024 instead of 512 pixels raised the test score by 0.011, while the score on caption–caption pairs, which contain no image and serve as a control, did not change.

3.2 Cross-encoders with citation-graph pretraining

A cross-encoder reads the two objects as one sequence of at most 512 tokens, so that each object can attend to the other, and predicts the four classes from the first token. We initialise from astroBERT (Grezes et al., 2021), whose vocabulary is built from astronomy text, and from SciNCL (Ostendorff et al., 2022), which is trained on citation neighbourhoods.

Before task fine-tuning, each cross-encoder is pretrained for one epoch on 600K pairs generated from the public corpus, 60K for each allowed combination of pair type and label. *same_figure* pairs join the OCR text of a figure with its own caption, *same_paper* pairs join two objects of one paper,

related_papers pairs join objects of two papers linked by a reference, and *unrelated_papers* pairs join objects of unlinked papers. Half of the *unrelated_papers* pairs are drawn from topically similar papers, so that similarity alone does not reveal the label. No cross-encoder is pretrained on a paper from the fold on which it is evaluated.

Pretraining, which targets generalisation to unseen papers, raised the test score by 0.022. It first imitated public image objects with *pseudo-OCR* text: captions with words dropped to resemble OCR output. Using real OCR text of the public figures instead added a further 0.038 to the cross-validated macro-F1 of a single cross-encoder, all of it in the two pair types with images.

The final *ensemble* averages 15 cross-encoders that differ in backbone, pretraining data, OCR input (whole image or tiles), and maximum length (256 or 512 tokens). The seven members pretrained on real OCR text get half of the total weight, the eight others the rest.

3.3 Stacker

The *stacker* is a small second-level classifier, one per pair type, that combines the ensemble probabilities with similarity and rank features from seven frozen embedding spaces, such as SigLIP 2 and SciNCL (Appendix H). It adds 0.004 macro-F1 over the ensemble on held-out data, a gain within noise (§4).

3.4 LLM classifiers and blending

The error analysis of the cross-encoder system found no further cue to add (§5), and adding more cross-encoders of the same kind had begun to lower the cross-validated score. We therefore added models of a different kind that read the same text input. Qwen3-8B (Yang et al., 2025) reads both objects and predicts the four classes through a new classification head; we train LoRA adapters (Hu et al., 2022) and the head for two epochs (Appendix A), without citation-graph pretraining. Gemma-4-12B (Gemma Team, 2026), from another model family, is trained the same way. These LLMs differ from the cross-encoders in size, pretraining data, and architecture at once, so we describe their contribution as making different errors, not adding capacity (Appendix G).

The final label is the argmax of the equal-weight average of the stacker and two LLM probabilities, with *same_figure* disallowed outside image–caption pairs.

Table 1: Selected development steps; each row adds to the row above (all steps in Appendix B). Below the rule: blends with stacker:Qwen:Gemma weights. CV: five-fold cross-validated macro-F1 (†: fold 1 only, where the row above scores 0.6991); Test: macro-F1 on the full test set. Shaded: gain above the test noise threshold of 0.014.

System	CV ↑	Test ↑
Frozen embeddings, rank features	0.5087	0.4972
+ OCR text features	0.5584	0.5557
+ symmetric cross-encoder	0.5908	0.5810
+ citation-graph pretraining	0.5979	0.6030
+ OCR at 1024 px, minor changes	0.6100	0.6193
+ astroBERT and SciNCL ensemble	0.6585	0.6695
+ 512-token inputs	0.6697	0.6795
+ real OCR text in pretraining	0.6976	0.7058
+ learned image projection features	0.7029	0.7072
+ Qwen3-8B (1:1)	0.7153†	0.7329
+ Gemma-4-12B (2:1:1)	–	0.7397
+ Gemma-4-12B (1:1:1, final)	–	**0.7421**

4 Experiments

Setup. We use five-fold cross-validation stratified by pair type and label, with the same folds for every component. *Held-out* predictions come from the model that did not train on the pair's fold, and CV is the macro-F1 over all of them. Unlike the test set, our folds share papers, so CV undervalues some changes (Appendix D). Qwen3-8B was validated on one fold only (fold 1, 2,000 pairs) and Gemma-4-12B not at all; each submitted LLM was trained once on all 10,000 pairs. A bootstrap gives the *noise threshold*: at the 95% level, two test scores must differ by 0.014 to be distinguishable, and two configurations compared on the same held-out pairs by about 0.008 (Appendix D). Table 1 traces the development of the system; Appendix C lists changes that did not help.

LLM blend. On fold 1, Qwen3-8B alone scores 0.6759, below the cross-encoder system (0.6991). Yet averaging the two with equal weights raises macro-F1 to 0.7153 (+0.016, paired 95% CI $[+0.002, +0.030]$) and the test score from 0.7072 to 0.7329. This is the last step supported by held-out data. Gemma-4-12B, which agrees with Qwen3-8B on only 79.4% of test labels, had no held-out predictions, so we submitted two weightings fixed in advance (Table 1) and report the higher score. That choice was made on the test set. It is worth 0.002 and the second LLM at most 0.009, both below the 0.014 threshold, so neither is established (Appendix D).

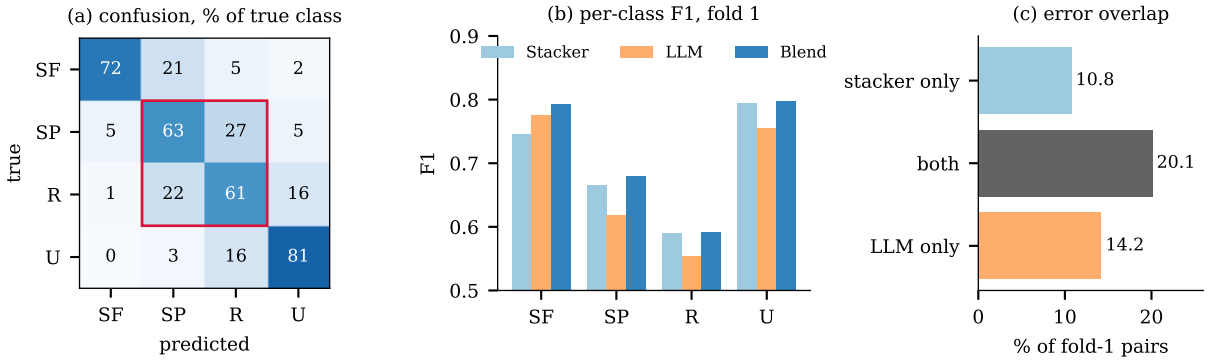


Figure 2: (a) Confusion matrix of the cross-encoder ensemble on held-out predictions for all 10,000 training pairs (% of true class; SF, SP, R, U abbreviate the labels). (b) Per-class F1 on fold 1. (c) Fold-1 pairs misclassified by one model or by both. "Stacker" is the cross-encoder system; "LLM" is Qwen3-8B.

5 Analysis

The dominant error of the first stage is *same_paper* versus *related_papers*. In Figure 2a, 27% of *same_paper* pairs are predicted as *related_papers* and 22% of *related_papers* pairs as *same_paper*; together these errors affect 14.7% of all pairs, against 9.9% for *related_papers* versus *unrelated_papers*. They are equally common in caption–caption pairs, which contain no image, so they are not a vision problem.

The errors follow similarity. We grouped the caption–caption pairs of these two classes by the word overlap of their captions, a proxy for topical similarity. From the least to the most similar group, the share of *same_paper* pairs rises from 32% to 90%, so similarity is valid evidence. The error rate falls from 42% to 10%. Errors thus occur on the pairs where this evidence misleads: *same_paper* pairs that look different and *related_papers* pairs that look alike.

No further cue we tested separates the classes. For separating *same_paper* from *related_papers*, shared object identifiers reach an area under the ROC curve (AUC) of 0.56, cross-references such as "same as Fig. N" 0.50, self-references such as "our sample" 0.50, and author–year citations 0.48. Chance is 0.50, and the cross-encoder reaches 0.85. A contrastive model trained with negatives from citing papers did remove topic from its embedding (correlation with topical similarity fell from 0.43 to 0.24), yet its *same_paper*-versus-*related_papers* AUC dropped by 0.01. Topical similarity is itself the main evidence for this distinction, so removing it discards signal (Appendix F).

The LLM makes partly different errors. On fold 1, 20.1% of pairs are misclassified by both the cross-encoder system and Qwen3-8B, 10.8% by the cross-encoder system only, and 14.2% by Qwen3-8B only (Figure 2c). The errors are correlated: independent errors would coincide on only 10.6% of pairs. Still, 35% of the cross-encoder system's errors and 41% of Qwen3-8B's are not made by the other model, and these are the pairs that averaging can correct. The blend improves *same_figure* and *same_paper* most (Figure 2b), and *related_papers* remains the hardest class.

6 Conclusion

Our AstroCLIMB system reads figures as text, pretrains cross-encoders on citation-graph pairs, and blends them with two LoRA-tuned LLMs, reaching 0.7421 test macro-F1. Its most frequent error, telling a paper from the papers it cites, was reduced by averaging with LLMs that make different errors, not by further cues. Whether a vision–language LLM would beat OCR text remains open.

Limitations

The Qwen3-8B blend was validated on a single fold of 2,000 pairs, Gemma-4-12B was not validated, and the final weights were chosen between two test submissions, so the final score is slightly optimistic. The last result supported by held-out data is the Qwen3-8B blend (0.7329). Each LLM was trained once with one random seed. We compared OCR text only with image embeddings and patch tokens from encoders up to the size of SigLIP 2. A vision–language LLM that reads the raw figure was not tested and may change that conclusion. The LLMs differ from the cross-encoders in size, pretraining

data, and architecture at once, so our results do not show which of these makes their errors different. Because figures enter the model only as OCR text, figures with little or no text, such as images of the sky, contribute little signal. Each LLM requires a 40 GB GPU, which is more than the organisers expected participants to use. DINOv3 is distributed under a licence not approved by the Open Source Initiative and was used only to compute frozen research features.

References

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *Computer Vision – ECCV 2022*. Springer.

Bishwaranjan Bhattacharjee, Aashka Trivedi, Masayasu Muraoka, Muthukumaran Ramasubramanian, Takuma Udagawa, Iksha Gurung, Nishan Pantha, Rong Zhang, Bharath Dandala, Rahul Ramachandran, Manil Maskey, Kaylin Bugbee, Michael M. Little, Elizabeth Fancher, Irina Gerasimov, Armin Mehrabian, Lauren Sanders, Sylvain V. Costes, Sergi Blanco-Cuaresma, and 17 others. 2024. INDUS: Effective and efficient language models for scientific applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 98–112, Miami, Florida, US. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Felix Grezes, Thomas Allen, Tirthankar Ghosal, and Sergi Blanco-Cuaresma. 2023. Function of citation in astrophysics literature (FOCAL): Findings of the shared task. In *Proceedings of the Second Workshop on Information Extraction from Scientific Publications*, pages 143–147, Bali, Indonesia. Association for Computational Linguistics.

Felix Grezes, Jennifer Lynn Bartlett, Kelly Lockhart, Alberto Accomazzi, Ethan Seefried, Anjali Pandiri, and Tirthankar Ghosal. 2025. Overview of TRACS: the telescope reference and astronomy categorization dataset & shared task. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, pages 5–15, Mumbai, India and virtual. Association for Computational Linguistics.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J. Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S. Grant, Donna M. Thompson, Roman Chyla, Stephen McDonald, Timothy W. Hostetler, Matthew R. Templeton, Kelly E. Lockhart, Nemanja Martinovic, Shinyi Chen, Chris Tanner, and Pavlos Protopapas. 2021. Building astroBERT, a language model for Astronomy & Astrophysics. *Preprint*, arXiv:2112.00590.

Felix Grezes, Sergi Blanco-Cuaresma, Thomas Allen, and Tirthankar Ghosal. 2022. Overview of the first shared task on detecting entities in the astrophysics literature (DEAL). In *Proceedings of the First Workshop on Information Extraction from Scientific Publications*, pages 1–7, Online. Association for Computational Linguistics.

Felix Grezes, Kelly Lockhart, Alberto Accomazzi, Atilla Alkan, Ethan Seefried, and Tirthankar Ghosal. 2026. AstroCLIMB: Astronomy citation linking from illustrations - a multimodal benchmark (overview of the WASP 2026 shared task). In *Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao Huang. 2021. SciCap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3258–3264, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

Mindee. 2021. docTR: Document text recognition. GitHub repository.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, and 7 others. 2025. DINOv3. *Preprint*, arXiv:2508.10104.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Hyperparameters

Compute. On one A100 GPU, a pretraining epoch takes about 12 minutes; training and test prediction take about 110 minutes for each LLM.

Citation-graph pretraining. One epoch, batch size 64, learning rate 3×10^{-5}, AdamW with weight decay 0.01, maximum length 192 tokens, 600K pairs. In pretraining and in task fine-tuning, the two objects are swapped with probability 0.5.

LLM classifier. Qwen3-8B in bfloat16, LoRA rank 16, $\alpha = 32$, dropout 0.05 on the query, key, value, and output projections; classification head trained in float32; two epochs, learning rate 10^{-4} with 5% linear warm-up, batch size 8 with two accumulation steps, gradient clipping at 1.0, maximum length 768 tokens; the two objects are presented as "Object 1 — … Object 2 — …" and swapped with probability 0.5 during training. Gemma-4-12B uses the same settings plus a linear head on its text backbone's last-token hidden state and training batches grouped by length, which cuts padding from 322 to 6 tokens per sequence.

B All development steps

Table 2 lists every submission that changed the system, in order; Table 1 merges some of these rows. Rows 18–22 have no five-fold score: row 18 changes test predictions only, Qwen3-8B was validated on fold 1 only (0.7130 for row 19 and 0.7153 for row 20), and Gemma-4-12B was not validated. Rows 6 and 18 were dropped, so rows 7 and 19 build on rows 5 and 17.

Table 2: All development steps. CV: five-fold cross-validated macro-F1; Test: macro-F1 on the full test set. Blend weights in rows 19–22 are given in Table 1.

#	Change	CV ↑	Test ↑
1	frozen embeddings, rank features	0.5087	0.4972
2	+ principal components	0.5140	0.5031
3	+ OCR text features	0.5584	0.5557
4	+ cross-encoder	0.5818	0.5671
5	+ both input orders, shared folds	0.5908	0.5810
6	+ image patches, paper clusters	0.5945	<0.5810
7	row 5 + graph pretraining	0.5979	0.6030
8	+ 6.6× citation edges	0.6047	0.6067
9	+ OCR at 1024 px	0.6088	0.6173
10	+ PNG colour-profile match	0.6100	0.6193
11	+ SciNCL/astroBERT ensemble	0.6401	0.6528
12	+ pretrained astroBERT	0.6585	0.6695
13	+ 512-token inputs	0.6697	0.6795
14	+ real OCR text in pretraining	0.6933	0.7021
15	+ 7 real-OCR members	0.6976	0.7058
16	+ image projection features	0.7015	0.7072
17	contrastive projection loss	0.7029	0.7072
18	+ transitivity repair	–	0.7067
19	row 17 ensemble + Qwen3-8B, 1:1	–	0.7292
20	row 17 stacker + Qwen3-8B, 1:1	–	0.7329
21	+ Gemma-4-12B, weights 2:1:1	–	0.7397
22	+ Gemma-4-12B, weights 1:1:1	–	0.7421

C Changes that did not help

Table 3 lists the changes we tested that did not improve the system. Each was measured against the same system without the change; details of the image inputs and of the contrastive objective are in Appendices E and F.

Table 3: Changes that did not help. Δ: change in macro-F1 relative to the same system without the change, cross-validated unless marked Test (full test set). The last row is an absolute score.

Change	Δ
Image patches in the cross-encoder (best of 4)	-0.024
Image embeddings as extra input tokens	-0.001
INDUS instead of astroBERT	-0.103
6.6× more citation edges in pretraining	$+0.003$
Self-training on test pseudo-labels	0.000
Transitivity repair on test triangles	Test -0.001
Clustering objects into paper profiles	Test < 0
Zero-shot Qwen3-8B as the classifier	0.378

D Evaluation protocol and use of the test set

Noise threshold. We have no test labels, so we estimated the sampling noise of macro-F1 from the held-out predictions of the cross-encoder ensemble for the 10,000 training pairs, a set of the same size as the test set. Resampling these pairs with replacement 2,000 times gives a standard deviation of 0.0052 for one score. The difference between two independently sampled scores then has a standard deviation of $\sqrt{2} \times 0.0052$, so two test scores must differ by $1.96 \times \sqrt{2} \times 0.0052 \approx 0.014$ to be distinguishable at the 95% level. This threshold is conservative for two systems scored on the same test pairs, because part of the sampling noise is shared, but the paired version cannot be computed without test labels. On held-out predictions it can: resampling the same pairs for both configurations gives a threshold of about 0.008 on 10,000 pairs. On fold 1 alone (2,000 pairs), intervals are wider, such as $[+0.002, +0.030]$ for the Qwen3-8B blend.

How the test set was used. The leaderboard is computed on the full test set, so every test score in Tables 1 and 2 was seen during development. These scores entered our decisions in three ways. First, changes were developed by cross-validation and then submitted, and we treated a cross-validated gain below 0.005 as unconfirmed until the test score agreed. Second, two submitted changes were dropped because the test score fell (rows 6 and 18 of Table 2). Third, the final blend weights were chosen between two weightings fixed in advance (rows 21 and 22). The weight of the Qwen3-8B blend was not tuned on the test set: equal weights were validated on fold 1. In addition, the rank features of the stacker (Appendix H) are computed among the objects of the test set itself, which uses test inputs but no test labels.

The final blend weights. Gemma-4-12B was trained only once, on all 10,000 pairs, so no held-out predictions existed for choosing its weight. We therefore fixed two weightings of the stacker, Qwen3-8B, and Gemma-4-12B before submitting either. The first, 2:1:1, treats the two LLMs as one group and keeps the 1:1 balance validated on fold 1; the second is 1:1:1. They scored 0.7397 and 0.7421, and we report the higher one. The difference of 0.002 is far below the 0.014 threshold, so the test set gives no evidence that one weighting is better. The gain of Gemma-4-12B over the Qwen3-8B blend (0.007 or 0.009) is also below the threshold. The last result supported by held-out data is therefore the Qwen3-8B blend, with a test score of 0.7329.

Why cross-validation and test scores disagree. Our cross-validation differs from the test set in two ways. First, the folds are split by pair, so a validation fold shares papers with its training folds, whereas test papers are new. A change that improves generalisation to unseen papers is therefore undervalued. Second, a test object appears in 1.79 pairs on average against 1.14 in training, so a better representation of a single object, such as better OCR, is rewarded more often on the test set. The steps we can compare are consistent with this. Citation-graph pretraining gained 3.3 times more on the test set than in cross-validation (row 7 of Table 2) and higher OCR resolution 2.6 times more (row 9), whereas adding pretraining pairs of the same kind, which changes neither property, gained only 0.55 times as much (row 8). Each ratio rests on single submissions whose differences are close to the noise threshold, so we offer this as an observation and not as a tested claim. Splitting folds by paper would remove the first difference, and we suggest it for future editions.

E Image inputs compared with OCR text

Table 4 lists every way we gave a model the image itself. All rows were measured on an early version of the system, before citation-graph pretraining, with one model per row. In a bi-encoder, each object is encoded separately into one vector by a SigLIP 2 tower, and a small classifier compares the two vectors. In the cross-encoder rows, 16 patch tokens from a frozen image encoder pass through a learned linear layer and enter the input sequence of the text model.

Table 4: Inputs for image objects on an early version of the system. CV: five-fold cross-validated macro-F1 of a single model.

Input for an image object	CV ↑
Cross-encoder	
OCR text	**0.556**
OCR text + MatCha patches	0.533
OCR text + SigLIP 2 patches	0.532
SigLIP 2 patches only	0.37
Bi-encoder	
Pooled SigLIP 2 embedding, frozen tower	0.353
Pooled SigLIP 2 embedding, fine-tuned tower	0.312

One reason for the gap in the first experiment was that the layer mapping patch tokens into the text model was trained from scratch on 8,000 pairs. We repeated the test on a final cross-encoder (0.682) with that layer pretrained on 554K public pairs that contain images. SigLIP 2 embeddings as extra input tokens changed the score by -0.001, and visual similarity written into the input as words by -0.002, both within noise.

These comparisons cover image encoders up to the size of SigLIP 2 attached to a BERT-sized text model. We did not test a vision–language LLM that reads the raw figure, such as a Qwen-VL model, which is the natural counterpart of OCR text fed to an LLM. Our finding is therefore that OCR text was the best input among those in Table 4, not that OCR is better than image models in general.

F Contrastive learning in this task

Deciding whether two objects share a paper resembles contrastive representation learning, and contrastive methods enter our system in several places. This appendix collects what each of them contributed.

Contrastively pretrained encoders as features. SigLIP 2 is trained with an image–text contrastive loss and SciNCL with a contrastive loss over citation neighbourhoods. Cosine similarities and rank features in these and five other frozen embedding spaces give 0.509 cross-validated macro-F1 (row 1 of Table 2). Comparing two separately computed embeddings is thus a weak classifier here, and reading the pair jointly with a cross-encoder is what raised the score (rows 4 and 5). As a single cross-encoder, pretrained SciNCL scores below astroBERT (0.583 against 0.620), but averaging the two gained 0.034 on the test set (row 11).

Fine-tuning a contrastive encoder. Updating the SigLIP 2 towers of a bi-encoder with the classification loss on 8,000 pairs lowered the score from 0.353 to 0.312 (Table 4), so we kept all contrastive encoders frozen.

A contrastive projection for image–image pairs. On frozen SigLIP 2 embeddings of public figures, we trained a projection that brings figures of the same paper together and used its cosine similarity as a stacker feature (§3.3). Trained with InfoNCE it gave 0.7029 cross-validated macro-F1, against 0.7015 with a binary cross-entropy loss and 0.6976 without the feature (rows 15 to 17); these differences are within noise. A projection trained directly to separate citing from unrelated papers was worse at that very distinction (AUC 0.62) than the same-paper projection (AUC 0.67).

Hard negatives in pretraining. Half of the *unrelated_papers* pretraining pairs are drawn from topically similar papers (§3.2), the cross-encoder counterpart of hard-negative sampling.

A contrastive objective aimed at the dominant error. With random negatives, a contrastive model can succeed by encoding topic alone. To make it encode which paper an object belongs to, we drew the negatives of each anchor only from papers linked to the anchor's paper by a citation. The model was a two-layer projection to 256 dimensions on frozen SigLIP 2 embeddings, trained with InfoNCE on 36,589 public figures and their captions from papers outside the training set. The positive was another object of the anchor's paper, and each anchor had 16 negatives. The control used the same anchors, architecture, and loss, with negatives from random other papers. Before training we fixed what success would look like: a lower correlation with topical similarity (the cosine of INDUS embeddings), a lower *same_paper*-versus-*unrelated_papers* AUC because topic no longer helps, and a higher *same_paper*-versus-*related_papers* AUC.

Table 5: Effect of citation-linked negatives relative to random negatives, measured on training pairs. SP, R, U abbreviate *same_paper*, *related_papers*, *unrelated_papers*. The topic correlation was measured on caption–caption pairs only.

	image–image	caption–caption
Correlation with topic	–	$0.43 \rightarrow 0.24$
AUC, SP vs. U	-0.108	-0.121
AUC, SP vs. R	-0.011	-0.010

Table 5 shows that the first two predictions held and the third did not. The objective removed topic, but topical similarity is itself the main evidence that two objects share a paper and not merely a citation link (median cosine similarity 0.54 for *same_paper* pairs against 0.42 for *related_papers* pairs). Removing it thus discarded signal without exposing another one. On caption–caption pairs the training loss with citation-linked negatives also fell to 0.13, against 0.73 with random negatives, suggesting that the model memorised the few citation neighbours of each paper.

G What the LLMs add

An LLM differs from our cross-encoders in at least three ways at once: size (8 and 12 billion against 110 million parameters), pretraining data (general-domain text against astronomy text or citation neighbourhoods), and architecture (a decoder reading up to 768 tokens against an encoder reading up to 512). We ran no experiment that separates these factors, so we do not attribute the gain of the blend to size. What we measured is the following.

Alone, the fine-tuned LLM is not stronger. Qwen3-8B scores 0.676 macro-F1 on fold 1, against 0.699 for the cross-encoder system.

Its errors differ. Of Qwen3-8B's errors, 41% are not made by the cross-encoder system, and 35% in the other direction (§5); averaging gains 0.016. The two models also differ by class and pair type: Qwen3-8B is better than the cross-encoder system on *same_figure* (F1 0.775 against 0.746) but much weaker on image–image pairs (macro-F1 0.586 against 0.679), where OCR text is often little more than axis numbers.

Pretraining knowledge does not help without fine-tuning. Zero-shot, with the four labels scored in one forward pass, Qwen3-8B reaches 0.378 macro-F1 and predicts *related_papers* for 2 of the 10,000 training pairs, so it does not recall which papers cite which. In a zero-shot binary probe on *same_paper* and *related_papers* caption–caption pairs, it reaches an AUC of 0.63 on 400 pairs from the less similar half (cross-encoder: 0.78) and 0.74 on 148 pairs from the most similar groups (cross-encoder: 0.86). Combining its ranks with those of the cross-encoder lowers the AUC on the less similar pairs (95% CI of the change $[-0.070, -0.010]$). The zero-shot LLM thus follows the same similarity as the cross-encoder and is weakest on the same pairs.

These results argue against memorised citation links as the source of the gain. They leave open whether fine-tuning draws on pretraining knowledge in ways a zero-shot test cannot reveal, and they do not separate size from architecture.

H Stacker features

For every pair we compute cosine similarities in seven frozen embedding spaces, among them SigLIP 2, DINOv3 (Siméoni et al., 2025), SciNCL, and INDUS (Bhattacharjee et al., 2024). Rank features record how close each object is to the other relative to all objects of the same split, using cross-domain similarity local scaling (CSLS; Conneau et al., 2018) to discount objects that are close to many others. Image–image pairs also get a SigLIP 2 similarity after a projection trained on the public corpus to detect figures of the same paper (Appendix F). For each pair type, a classifier chosen by cross-validation (logistic regression, a two-stage variant, or a small multilayer perceptron) maps these features and the ensemble probabilities to the four classes. On held-out data the stacker adds 0.004 macro-F1 over the ensemble, which is within noise.

I Related work

Prior work on scientific figures has mostly generated captions (Hsu et al., 2021); AstroCLIMB instead asks how figures, captions, and papers relate. Scientific language models learn from domain text (Grezes et al., 2021; Bhattacharjee et al., 2024) or from citation neighbourhoods (Cohan et al., 2020; Ostendorff et al., 2022); we use both and turn the citation graph into labelled pairs.